Improving methodological robustness in cross-cultural organizational research

van de Vijver, F.J.R.; Fischer, R.

Published in:
Handbook of culture, organizations, and work

Publication date:
2009

Document Version
Publisher's PDF, also known as Version of record

PART IV

Future Directions in Theory and Research

CHAPTER 18

Improving methodological robustness in cross-cultural organizational research

FONS J. R. VAN DE VIJVER and RONALD C. FISCHER

Some of the largest and best known cross-cultural psychological projects come from the domain of organizational research; good examples are Hofstede's (1980, 2001) study on attitudes of IBM employees and the GLOBE study which involved sixty-two countries (House *et al.* 2003). However, these large projects are somewhat atypical in that most cross-cultural organizational studies involve two or three cultures. The current chapter provides an overview of basic issues in cross-cultural organizational research. The combination of a large interest in cross-cultural organizational research and the lack of a formal training of many researchers in cross-cultural methods create the need to reflect on these basic issues. The central question is how we can improve the methodological robustness of our research which, as we expect, will contribute to the validity and replicability of the conclusions derived from our research. We do not discuss the theories that are used in this field but focus on the methodological issues that are common to cross-cultural research (a good overview of current theories can be found in Smith, Bond, and Kagitcibasi, 2006).

The chapter deals with two kinds of methodological issues. The first involves the basic question of the comparability of constructs and scores across cultures (Poortinga, 1989). Comparability of scores across individuals obtained in a monocultural setting is typically taken for granted. We readily compare scores from participants in different organizations once we have established an adequate reliability and factorial composition of the instrument. Managers routinely use survey instruments and tests developed in different cultural contexts to make decisions about selecting or promoting employees, to judge morale and satisfaction of staff or to evaluate effectiveness of training programmes, interventions or organizational effectiveness. However, the implicit assumption of comparability cannot be taken for granted in cross-cultural research. Comparability can be challenged in various ways. For example, cross-cultural differences in views on controversial topics such as abortion and soft-drug use may be influenced by differences in national laws, the societal climate of (in)tolerance surrounding these topics, and ensuing differences in social desirability. Our chapter primarily focuses on these factors in the context of cross-cultural applications of standard instruments or tests.

The second issue discussed in this chapter involves the multilevel design of cross-cultural organizational studies. Models have been developed in the last decades to account for the complex data structure of such studies which involve participants nested in organizations nested in cultures (Dansereau, Alutto and Yammarimo, 1984; Raudenbush and Bryk, 2001; Muthén, 1991, 1994). Cross-cultural psychological studies often draw inferences on cultures on the basis of individual-level scores. Multilevel analyses therefore need to address the following questions:

(a) What is the most appropriate level of analysis (individual, group, organization, industry, national culture, etc.)?
(b) Do concepts that exist at more than one level have the same meaning at all levels (isomorphism across levels)?
(c) What is the linkage of constructs across levels (e.g., influence of higher-level constructs on lower-level constructs)?

The first question needs to be addressed theoretically as well as methodologically. Researchers need to specify their appropriate level of theory and then measure the variables at this level. Much cross-cultural research uses aggregated scores. Any statistical test of differences in means, such as a *t* test or analysis of variance, assumes that the meaning of scores does not change after aggregation. We assume that the mean score is a good reflection of the standing of the culture on the underlying construct. Techniques are available to address what level is empirically justified. We also know that scored aggregation can lead to a change of meaning. Additional constructs can influence country-level differences. The statistical models that have been developed can address the question to what extent scores that are aggregated still have the same meaning after aggregation. For example, do scores on leadership preference still reflect this construct after scores have been aggregated at country level or are country-level differences influenced by additional constructs such as social desirability? Finally, we can investigate the relationships across levels. The most common question that can be statistically addressed refers to the prediction of a psychological variable (e.g., leadership preferences) by means of individual-level variables (e.g., education), organizational-level variables (e.g., size), and country-level variables (e.g., power distance and Gross National Product).

The first section of the chapter deals with scoring comparability; a taxonomy of bias and equivalence is presented that allows us to systematically describe levels of comparability. The second section deals with multilevel issues. Conclusions are drawn in the final section.

Bias and equivalence

An important question to consider in the initial stages of a project involves the choice of instrument. There are essentially three options: use an existing instrument; adapt an existing instrument; or develop a new instrument (Van de Vijver, 2003). Even in a project in which an existing (usually western) instrument has to be used, it is still important to consider the appropriateness of the existing instrument in the target culture. Appropriateness depends on linguistic, cultural, and psychometric criteria. Linguistic criteria involve the denotative and connotative meaning of stimuli and their comprehensibility. Cultural criteria involve the compliance with local norms and habits. Psychometric criteria involve characteristics involve the common criteria of validity and reliability.

The first option, called adoption, amounts to a close translation of an instrument in a target language. This option is the most frequently chosen in empirical research because it is simple to implement, cheap, has a high face validity, and retains the opportunity to compare scores obtained with the instrument across all translations. The aim of these translations often is the comparison of averages obtained in different cultures (does culture A score higher on construct X than does culture B?). Close translations have an important limitation: they can only be used when the items in the source and target language versions have an adequate coverage of the construct measured and no items show bias. Standard statistical techniques for assessing equivalence (e.g., Van de Vijver and Leung, 1997) should be applied to assess the similarity of constructs measured by the various language versions. However, even when the structures are identical, there is no guarantee that the translations are all culturally viable and that a locally developed instrument would cover the same aspects.

The second (and increasingly popular) option is labeled adaptation. It usually amounts to the close translation of some stimuli that are assumed to be adequate in the target culture, and to a change of other stimuli when a close translation would lead to linguistically, culturally or psychometrically inappropriate measurement (e.g., a questionnaire has the item "invite your boss over for a birthday party at your house") to express the idea of emotional closeness in organizations. However, the implicit assumption that birthday parties are a culturally important institution is not universally valid. A behavior could then be identified that comes close to the original in terms of psychological meaning (e.g., a meeting with a superior in an informal family setting).

The third option is called assembly. It involves the compilation of an entirely new instrument. It is

the preferable choice if a translation or adaptation process is unlikely to yield an instrument with satisfactory linguistic, cultural, and psychometric accuracy. An assembly will lead to an emic, culture-specific instrument. An assembly maximizes the cultural suitability of an instrument, but it will preclude any numerical comparisons of scores across cultures.

There is no single best option. The choice for either option should be based on various factors. If the aim is to compare scores obtained with an instrument in different cultures, a close translation is the easiest procedure. However, the cultural adequacy of the instrument in the target culture has to be demonstrated. The "quick and dirty" practice of preparing a close translation, administering it in a target culture, and comparing the scores in a *t* test without any concern for the cultural and psychometric adequacy of the measure is hard to defend. If the aim is to maximize the ecological validity of the instrument (i.e., to measure the construct in a target culture as adequate way), an adaptation or assembly is preferable. Culture-specific items can increase the validity of research findings in specific cultural contexts and give us a better contextual understanding of the psychological processes (Bhagat and McQuaid, 1982), but they also decrease the comparability of the findings across cultural groups. Statistical tools, such as item response theory and structural equation modeling, can deal with an incomplete overlap in indicators across cultures (Van de Vijver and Leung, 1997). However, if the number of culture-specific items is large, the comparability of the construct or of the remaining items may be problematic. The maximization of cross-cultural comparability and of local validity may be incompatible in such cases. In the remainder of the chapter, we will deal with issues which are especially important for adopted and adapted instruments.

Bias

Bias refers to the presence of nuisance factors that challenge the comparability of scores across cultural groups. If scores are biased, their psychological meaning is culture dependent and cultural differences in assessment outcome are to be accounted for, partly or completely, by auxiliary psychological constructs or measurement artifacts.

The occurrence of bias has a bearing on the comparability of scores across cultures. The measurement implications of bias for comparability are addressed in the concept of *equivalence* (see Johnson, 1998, for a review). Equivalence refers to the comparability of test scores obtained in different cultural groups. Obviously, bias and equivalence are related; it is sometimes argued that they are mirror concepts. Bias, in this view, is synonymous to nonequivalence; conversely, equivalence refers to the absence of bias. This is not the view adopted here because, in the presentation of cross-cultural research methodology, it is instructive to disentangle sources of bias and their implications for score comparability.

Bias and equivalence are not inherent characteristics of an instrument, but arise in the application of an instrument in at least two cultural groups and the comparison of scores, patterns or item values. Decisions on the presence or absence of equivalence should be empirically based. The need for such validation and verification should not be interpreted as blind empiricism and the impossibility of implementing preventive measures in a study to minimize bias and maximize equivalence. On the contrary, not all instruments are equally susceptible to bias. For example, more structured test administrations are less prone to bias influences than are less structured sessions (assuming that the test administrations are adequately tailored to the cultural context and the test administration is not based on western manuals that neglect local communication conventions). Analogously, comparisons of closely related groups will be less susceptible to bias than comparisons of groups with a widely different cultural background.

Identification of bias and verification of equivalence are core theoretical as well as methodological problems of cross-cultural survey research (Smith, Bond, and Kagitcibasi , 2006). The validity of any comparison critically depends on the solution of these two issues. Malpass (1977) pointed out that methodological problems in cross-cultural research are often theoretical problems in disguise. If we measure some construct in two or more samples, we need to understand any potential variable

Table 18.1 Sources of bias in cross-cultural assessment

Type of Bias	Source of Bias
Construct bias	• Only partial overlap in the definitions of the construct across cultures (e.g., filial piety, as described in the main text).
	• Differential appropriateness of the behaviors associated with the construct (e.g., items do not belong to the repertoire of one of the cultural groups).
	• Poor sampling of all relevant behaviors (e.g., short instruments are used to cover broad constructs).
	• Incomplete coverage of all relevant aspects/facets of the construct (e.g., not all relevant domains are sampled).
Method bias	Sample bias
	• Incomparability of samples (e.g., caused by differences in kinds of organizations, education, or motivation across cultures).
	• Differences in environmental administration conditions, physical (e.g., recording devices) or social (e.g., class size).
	• Ambiguous instructions for respondents and/or guidelines for administrators.
	Administration bias
	• Differential expertise of administrators/interviewers.
	• Tester/interviewer/observer effects (e.g., halo effects).
	• Communication problems between participant and interviewer (e.g., participant is not sufficiently proficient in language of testing).
	Instrument bias
	• Differential response styles (e.g., social desirability, extremity scoring, acquiescence).
	• Differential familiarity with stimulus material and/or response procedures (particularly relevant in cognitive testing).
Item bias	• Poor translation (e.g., linguistically equivalent translation of a word does not exist in source and target language).
	• Ambiguous items (e.g., double barreled items).
	• Nuisance factors (e.g., item may invoke additional traits or abilities).
	• Cultural specifics (e.g., incidental differences in connotative meaning and/or appropriateness of the item content).

that can have an impact on the scores in one of the samples. The central issue is that respondents may be responding to the researcher or administrator, the social context in which the research takes place and the specific task in other ways than we believe they are. It is important to understand the 'mind of the other' (Malpass, 1977), the meaning that is created by participants in different groups. The purpose of establishing equivalence is to examine this similarity in meaning. When we address the equivalence, we operationalize this similarity in meaning. For example, if the items of an instrument show similar associations with each other in different cultures, we argue that these items measure the same underlying constructs in these groups.

Sources of bias: construct, method, and item. In order to detect and/or prevent bias, we need to recognize what can lead to bias. Table 18.1 provides an overview of sources of bias, based on a classification by Van de Vijver and Tanzer (2004; cf. Van de Vijver and Poortinga 1997). Sources of bias are numerous, thus the overview is necessarily tentative.

Construct bias occurs when the construct measured is not identical across groups. Construct bias precludes the cross-cultural measurement of a construct with the same measure. Detection of construct bias requires some intimate familiarity with the culture being studied, which can be achieved by conducting local pilot studies in the initial stages of a project or using local insiders

(see below). Embretson (1983) coined the term *construct underrepresentation* to describe the situation where an instrument insufficiently represents all the domains and dimensions relevant for a given construct in a given culture. There is an important difference between our term *construct bias* and Embretson's term. Whereas construct underrepresentation is a problem of instruments measuring broad concepts with too few indicators which can usually be overcome by adding items relating to these domains/dimensions, construct bias can only be overcome by adding items relating to new domains/dimensions. Clearly, identification of construct bias calls for detailed culture-specific knowledge.

Cross-cultural differences in the concept of depression are one example. Another empirical example can be found in Ho's (1996) work on filial piety (defined as a psychological characteristic associated with being "a good son or daughter"). The Chinese conception, according to which children are expected to assume the role of caretaker of elderly parents, is broader than the western. An inventory of filial piety based on the Chinese conceptualization covers aspects unrelated to the concept among western subjects, whereas a western-based inventory will leave important Chinese aspects uncovered. In western-based organizational settings, commitment has been conceptualized as a three-componential model (Cohen, 2003; Meyer and Allen, 1991; Meyer *et al.* 2002), differentiating affective, continuance and normative forms of commitment. Affective commitment is the emotional attachment to organizations and characterized by a genuine want or desire to belong to the organization as well as congruence and identification with the norms, values and goals of the organization. Continuance commitment focuses on the alleged costs associated with leaving or altering one's involvement with the organization, implying a perceived need to stay. Normative commitment is considered as a feeling of obligation to remain with the organization, capturing normative pressures and perceived obligations by important others.

The extent to which such definitions capture the understanding of commitment in different cultural contexts is yet unclear (Fischer and Mansell, 2008; Wasti and Oender, 2008). A meta-analysis by Fischer and Mansell (2008) showed that the three components showed considerable, but incomplete overlap in lower income contexts indicating that the components might not be functionally equivalent across economic contexts. Wasti (2002) argued that continuance commitment in a Turkish context is too narrowly defined. In more collectivistic contexts, loyalty and trust are important and strongly associated with paternalistic management practices. Therefore, employers are more likely to give trusted jobs to family members or friends, involving these individuals into relationships of dependency and obligation. This practice, in turn, leads to efforts to maintain "face" and one's credibility and attempts to return the favor. These normative pressures therefore become part of continuance commitment, involving both financial and rational considerations (investments, benefits as found in western contexts) as well as social costs (loss of face and credibility).

Yang and Bond (1990) presented indigenous Chinese personality descriptors and a set of American descriptors to a group of Taiwanese subjects. Factor analyses showed differences in the Chinese and American factor structures. Similarly, Cheung *et al.* (1996) found that the western-based five-factor model of personality (McCrae and Costa 1997) does not cover all the aspects deemed relevant by the Chinese to describe personality. In addition to the western factors of *extraversion*, *agreeableness*, *conscientiousness*, *neuroticism* (emotional stability), and *openness*, two further factors were found relevant for the Chinese context: *face* and *harmony*.

Construct bias can also be caused by differential appropriateness of the behaviors associated with the construct in the different cultures. An example of this comes from research on intelligence. Western intelligence tests tend to focus on reasoning and logical thinking (e.g., Raven's Progressive Matrices), while omnibus tests also contain subtests that tap into acquired knowledge (e.g., vocabulary scales for the Wechsler scales). When western respondents are asked which characteristics they associate with an intelligent person, skilled reasoning and extensive knowledge are frequently mentioned, as well as social aspects of intelligence. These social aspects are even more

prominent in everyday conceptions of intelligence in non-western groups. Kokwet mothers (Kenya) expect that intelligent children know their place in the family and the fitting behaviors for children, such as proper forms of address. An intelligent child is obedient and does not create problems (Segall *et al.* 1990).

Construct bias is also apparent in commitment research. Since Cole's (1979) initial comparison of behavioral commitment levels in Japan and the US, there has been a great interest in differences and similarities in commitment across cultural groups. However, researchers soon found out that high levels of behavioral commitment among Japanese workers (indicated by low turnover) were not strongly correlated with attitudinal commitment, as was found in the US. Therefore, the behavior of (or thoughts about) leaving one's organization was a good indicator of attitudinal commitment in the US, but not in Japan (for reviews, see Besser, 1993; Lincoln and Kalleberg, 1990; Smith, Fischer and Sale, 2001).

An important type of bias, called *method bias*, can result from such factors as sample incomparability, instrument differences, tester and interviewer effects, and the mode of administration. Method bias is used here as a label for all sources of bias emanating from factors often described in the methods section of empirical papers or study documentations. They range from differential stimulus familiarity in mental testing to differential social desirability in personality and survey research. Identification of methods bias requires detailed and explicit documentation of all the procedural steps in a study.

Among the various types of method bias, sample bias is more likely to increase with cultural distance. Recurrent rival explanations (which become more salient with a larger cultural distance) are cross-cultural differences in social desirability and stimulus familiarity (testwiseness). The main problem with both social desirability and testwiseness is their relationship with country affluence; more affluent countries tend to show lower scores on social desirability (see Chapter 13). Subject recruitment procedures are another source of sample bias in cognitive tests. For instance, the motivation to display one's attitudes or abilities may depend on the amount of previous exposure to psychological tests, the freedom to participate or not, and other sources that may show cross-cultural variation.

Administration bias can be caused by differences in the procedures or mode used to administer an instrument. For example, when interviews are held in respondents' homes, physical conditions (e.g., ambient noise, presence of others) are difficult to control. Respondents are more prepared to answer sensitive questions in self-completion contexts than in the shared discourse of an interview. Examples of social environmental conditions are individual (versus group) administration, the physical space between respondents (in group testing), or class size (in educational settings). Other sources of administration that can lead to method bias are ambiguity in the questionnaire instructions and/or guidelines or a differential application of these instructions (e.g., which answers to open questions are considered to be ambiguous and require follow-up questions). The effect of test administrator or interviewer presence on measurement outcomes has been empirically studied; regrettably, various studies apply inadequate designs and do not cross the cultures of testers and testees. In cognitive testing, the presence of the tester is usually not very obtrusive (Jensen, 1980). In survey research there is more evidence for interviewer effects (Singer and Presser, 1989). Deference to the interviewer has been reported; subjects were more likely to display positive attitudes to a particular cultural group when they are interviewed by someone from that group (e.g., Aquilino, 1994). A final source of administration bias is constituted by communication problems between the respondent and the tester/interviewer. For example, interventions by interpreters may influence the measurement outcome. Communication problems are not restricted to working with translators. Language problems may be a potent source of bias when, as is not uncommon in cross-cultural studies, an interview or test is administered in the second or third language of interviewers or respondents. Illustrations for such miscommunications between native and nonnative speakers can be found in Gass and Varonis (1991).

Instrument bias is a common problem in cognitive tests. An interesting example comes from

Piswanger's (1975) application of the Viennese Matrices Test (Formann and Piswanger, 1979). A Raven-like figural inductive reasoning test was administered to high-school students in Austria, Nigeria, and Togo (where the medium of instruction is Arabic). The most striking findings were cross-cultural differences in item difficulties related to identifying and applying rules in a horizontal direction (i.e., left to right). These differences were interpreted as bias due to the different directions in writing Latin and Arabic.

The third type of bias distinguished here refers to anomalies at item level and is called *item bias* or *differential item functioning*. According to a definition that is widely used in education and psychology, an item is biased if respondents with the same standing on the underlying construct (e.g., they are equally intelligent), but who come from different cultures, do not have the same mean score on the item. The score on the construct is usually derived from the total test score. Of all bias types, item bias has been the most extensively studied; various psychometric techniques are available to identify item bias (e.g., Camilli and Shepard, 1994; Van de Vijver and Leung, 1997). In a globalized working environment, the standardized application of uniform managerial and human resource practices requires that we test the applicability of test items for different populations. Item bias primarily applies to instruments where the same items are used to measure the construct in different samples. Including emic items that are non-comparable across groups can be informative for cultural purposes, but such items mostly preclude direct comparison.

Although item bias can arise in various ways, poor item translation, ambiguities in the original item, low familiarity/appropriateness of the item content in certain cultures, and the influence of cultural specifics such as nuisance factors or connotations associated with the item wording are the most common sources. For instance, if a geography test administered to pupils in Poland and Japan contains the item "What is the capital of Poland?," Polish pupils can be expected to show higher scores on the item than Japanese students, even if pupils with the same total test score were compared. The item is biased because it favors one cultural group across all test score levels. Even translations which are seemingly correct can produce problems. A good example is the test item "Where is a bird with webbed feet most likely to live?" which was part of a large international study of educational achievement (cf. Hambleton, 1994). Compared to the overall pattern, the item turned out to be unexpectedly easy in Sweden. An inspection of the translation revealed why: the Swedish translation of the English was "bird with swimming feet," which gives a strong clue to the solution not present in the English original.

How to deal with bias

The previous section contains real and fictitious examples of bias. It is important to note that bias can affect all stages of a project. Minimizing bias is thus not an exclusive concern of developers, administrators, or data analysts. Since bias can challenge all stages of a project, ensuring quality is a matter of combining good theory, questionnaire design, administration, and analysis. The present section presents various ways in which the types of bias discussed above can be dealt with.

A taxonomy of the main approaches to deal with bias is presented in table 18.2 (cf. Van de Vijver and Tanzer, 2004). Rather than attempting to provide an exhaustive taxonomy (which goes beyond the scope of the present chapter), an attempt is made to provide an overview of solutions that have been presented in the past and to suggest directions in which a possible solution may be found in the event that the table does not provide a ready-made answer.

It should be emphasized that the focus of this chapter is on comparative studies. Within this context, culture-specifics constitute a potential challenge to be overcome. This focus on similarity is sometimes seen as a focus on universal aspects and the denial of culture-specifics. We do not concur with this view as some of the most interesting cross-cultural differences may reside in the cultural specifics. Emic research which tries to understand the culture from within is very important and informative for organizational research (Bhagat and McQuaid, 1982). Knowledge of emic concepts is critical for conducting studies of that culture,

Table 18.2 Strategies for identifying and dealing with bias

Type of Bias	Strategies
Construct bias	• Decentering (i.e., simultaneously developing the same instrument in several cultures). • Convergence approach (i.e., independent within-culture development of instruments and subsequent cross-cultural administration of all instruments).
Construct bias and/or method bias	• Use of informants with expertise in local culture and language. • Use samples of bilingual subjects. • Use of local pilots (e.g., content analyses of free-response questions). • Nonstandard instrument administration (e.g., "thinking aloud"). • Cross-cultural comparison of nomological networks (e.g., convergent/discriminant validity studies, monotrait-multimethod studies. • Connotation of key phrases (e.g., examination of similarity of meaning of frequently employed terms such as "somewhat agree").
Method bias	• Extensive training of interviewers. • Detailed manual/protocol for administration, scoring, and interpretation. • Detailed instructions (e.g., with sufficient number of examples and/or exercises). • Use of subject and context variables (e.g., educational background). • Use of collateral information (e.g., test-taking behavior or test attitudes). • Assessment of response styles. • Use of test-retest, training and/or intervention studies.
Item bias	• Judgmental methods of item bias detection (e.g., linguistic and psychological analysis). • Psychometric methods of item bias detection (e.g., Differential Item Functioning analysis).

even if the study would be culture-comparative. However, from a methodological vantage point, cultural specifics need to be handled with care as, by definition, they are difficult or even impossible to compare across cultures. So, the focus on bias in comparative research is not meant to eliminate culture-specifics but to tell these apart from more universal aspects and to ascertain which aspects are universal and which are culture specific.

The first example of dealing with *construct bias* is cultural decentering (Werner and Campbell, 1970). A modified example can be found in the study of Tanzer, Gittler, and Ellis (1995). Starting with a set of German intelligence/aptitude tests, they developed an English version of the test battery. Based on the results of pilot tests in Austria and the US, both the German and English instructions and stimuli were modified before the main study was carried out. In the so-called convergence approach estimates are independently developed in different cultures and all instruments are then administered to subjects in all these cultures (Campbell, 1986).

A second set of remedies aims at a combination of construct and method bias. Another example is a large acculturation project, called ICSEY (International Comparative Study of Ethnic Youth). The project studies both migrant and host adolescents and their parents in thirteen countries, including migrants from about fifty different ethnic groups. Prior to the data collection, researchers met to decide on which instruments would be used. Issues like adequacy of the instrument vis-à-vis construct coverage and translatability (e.g., absence of colloquialisms and metaphorical expressions) were already factored into the instrument design, thereby presumably avoiding various possible problems in later stages. Other measures taken include using informants with expertise in local culture and language, samples of bilingual individuals, local pilots (e.g., content analyses of free-response questions), nonstandard instrument administration (e.g., thinking aloud), and a pretest study of the connotation of key phrases.

The cross-cultural comparison of nomological networks constitutes an interesting possibility to

examine construct and/or method bias. An advantage of this infrequently employed method is its broad applicability. The method is based on a comparison of the correlations of an instrument that may have indicators that vary considerably across countries with various other instruments. The adequacy of the instrument in each country is supported if it shows a pattern of positive, zero, and negative correlations that are expected on theoretical grounds. For example, views towards waste management, when measured with different items across countries, may have positive correlations with concern for the environment and air pollution and a zero correlation with religiosity. Nomological networks may also be different across cultures; Tanzer and Sim (1991) found, for example, that good students in Singapore worry more about their performance during tests than do weak students, whereas the contrary is commonly reported in many other countries. For the other components of test anxiety (i.e., tension, low confidence, and cognitive interference), no cross-cultural differences were found. The authors attributed the inverted worry-achievement relationship to characteristics of the Singaporean educational system, especially the "kiasu" (fear of losing out) syndrome, which is deeply entrenched in the Singaporean society, rather then to construct bias in the internal structure of test anxiety.

Various procedures have been developed that mainly address method bias. A first proposal involves the extensive training of administrators/interviewers. Such training and instructions are required in order to ensure that interviews are administered in the same way across cultural groups. If the cultures of the interviewer and the interviewee differ, as is common in studies involving multicultural groups, it is important to make the interviewers aware of the relevant cultural specifics such as taboo topics.

A related approach amounts to the development of a detailed manual and administration protocol. The manual should ideally specify the test or interview administration and describe contingency plans on how to intervene in common interview problems (e.g., specifying when and how follow-up questions should be asked in open questions).

The measures discussed attempt to reduce or eliminate unwanted cross-cultural differences in administration conditions so as to maximize the comparability of scores obtained. Additional measures are needed to deal with cross-cultural differences that cannot be controlled by careful selection and wording of questions or response alternatives. Education is a good example. Studies involving widely different groups cannot avoid that the samples studied differ substantially in educational background, which in turn may give rise to cross-cultural differences in scores obtained. In some studies it may be possible to match groups from different groups on education by sampling subjects from specified educational backgrounds. However, this approach can have serious limitations; the samples obtained may not be representative for their countries. This problem is particularly salient in comparisons of countries with a population with large differences in average educational level. For example, if samples of Canadian and South African adults are chosen that are matched on education, it is likely that at least one of the samples is not representative for its population. Clearly, if one is interested in a country comparison after controlling for education, this poor representativeness does not create a problem. If the two samples are obtained using some random sampling scheme, educational differences are likely to emerge. The question may then arise to what extent the educational differences can be held responsible for observed test score differences. For example, to what extent could differences in attitudes toward euthanasia be explained by educational differences? If individual-level data on education is available, various statistical techniques, such as covariance and regression analysis, can be used as to determine to what extent the observed country differences can be explained by educational differences (Poortinga and Van de Vijver, 1987). The use of such explanatory variables provides a valuable tool to examine the nature of cross-cultural score differences.

A perennial issue in survey research is the prevalence of response effects and styles, especially social desirability and acquiescence. Their role in cross-cultural research as a source of unwanted cross-cultural score differences should not be underestimated. For some of the response styles,

questionnaires are available; for example, the Eysenck Personality Questionnaire (Eysenck and Eysenck, 1975) has a social desirability subscale that has been applied in many countries. When response styles are suspected of differentially influencing responses as obtained in different cultural groups, the administration of a questionnaire to assess the response style can provide a valuable tool to interpret cross-cultural score differences.

There is empirical evidence indicating that countries differ in their usage of response scales. Hui and Triandis (1989) found that Hispanics tended to choose extremes on a five-point rating scale more often than white Americans, but that this difference disappeared when a ten-point scale was used. Similarly, Oakland, Gulek, and Glutting (1996) assessed test-taking behaviors among Turkish children, and their results, similar to those obtained with American children, showed that these behaviors are significantly correlated with the WISC-R IQ.

There are two kinds of procedures to assess item bias: judgmental procedures, either linguistic or psychological, and psychometric procedures. An example of a linguistic procedure can be found in Grill and Bartel (1977). They examined the Grammatic Closure subtest of the Illinois Test of Psycholinguistic Abilities for bias against speakers of nonstandard forms of English. In the first stage, potentially biased items were identified. Error responses of American black and white children indicated that more than half the errors on these items were accounted for by responses that are appropriate in nonstandard forms of English.

Equivalence

Four different types of equivalence are proposed here (cf. Van de Vijver and Leung, 1997; for a discussion of many concepts of equivalence, see Johnson, 1998). *Construct inequivalence* amounts to comparing *apples and oranges* without raising the level of comparison to that of *fruit* (e.g., the comparison of Chinese and western *filial piety*, discussed above). If constructs are inequivalent, comparisons lack a shared attribute, which precludes any comparison.

Structural or functional equivalence is found if an instrument administered in different cultural groups shows structural equivalence measures the same construct in all these groups. Structural equivalence has been addressed for various cognitive tests (Jensen 1980), Eysenck's personality questionnaire (Barrett *et al.* 1998), and the so-called five-factor model of personality (McCrae and Costa, 1997). Structural equivalence does not presuppose the use of identical instruments across cultures. A depression measure may be based on different indicators in different cultural groups and still show structural equivalence.

The third type of equivalence is called *measurement unit equivalence*. Instruments show this if their measurement scales have the same units of measurement, but a different origin (such as the Celsius and Kelvin scales in temperature measurement). This type of equivalence assumes interval- or ratio-level scores (with the same measurement units in each culture). Measurement unit equivalence applies when the same instrument has been administered in different cultures and a source of bias with a fairly uniform influence on the items of an instrument affects test scores in the different cultural groups in a differential way; for example, social desirability and stimulus familiarity influence scores more in some cultures than in others. When the relative contribution of both bias sources cannot be estimated, the interpretation of group comparisons of mean scores remains ambiguous.

At first sight, it may seem unnecessary or even counterproductive to define a level of equivalence with the same measurement units, but different origins. After all, if we apply the same interval-level scale in different groups, scores may be either fully comparable or, as in the case of nonequivalence, fully incomparable. The need for the concept of measurement unit equivalence may become clear by looking at the impact of differential social desirability or stimulus familiarity on cross-cultural score differences in more detail. Differential social desirability will create an offset in the scale in one of the cross-cultural groups: a score of, say, five in group A may be comparable to a score of nine in group B because of a higher social desirability in group B. Observed group differences in mean scores are then a mixture of valid cross-cultural

differences and measurement artifacts. A correction would be required to make the scores comparable (Fischer, 2004). It may be noted that the basic idea of score corrections needed to make scores fully comparable is also applied in covariance analysis, in which score comparisons are made after the disturbing role of concomitant factors (bias in the context of the present chapter) has been statistically controlled for.

Only in the case of *scalar (or full score) equivalence* can direct comparisons be made; this is the only type of equivalence that allows for the conclusion that average scores obtained in two cultures are different or equal. Scalar equivalence assumes the identical interval or ratio scales across cultural groups. It is often difficult to decide whether equivalence in a given case is scalar equivalence or measurement equivalence. For example, ethnic differences in intelligence test scores have been interpreted as due to valid differences (scalar equivalence) as well as reflecting measurement artifacts (measurement unit equivalence). Scalar equivalence assumes that the role of bias can be safely neglected. However, verification of scalar equivalence relies on inductive evidence. Thus it is easier to disprove scalar equivalence than to prove it (cf. Popper's falsification principle). Measuring presumably relevant sources of bias (such as stimulus familiarity or social desirability) and showing that they cannot statistically explain observed cross-cultural differences in a multiple regression or covariance analysis is an example of falsifying a rival hypothesis.

Structural, measurement unit, and scalar equivalence are hierarchically ordered. The third presupposes the second, which presupposes the first. As a consequence, higher levels of equivalence are more difficult to establish. It is easier to verify that an instrument measures the same construct in different cultural groups (structural equivalence) than to identify numerical comparability across cultures (scalar equivalence). But one should bear in mind that higher levels of equivalence allow for more detailed comparisons of scores across cultures. Whereas only factor structures and nomological networks can be compared in the case of structural equivalence, measurement unit and full score or scalar equivalence allow for more fine grained analyses of cross-cultural similarities and differences, such as comparisons of mean scores across cultures in *t* tests and analyses of (co)variance.

The use of exploratory and confirmatory factor analysis in establishing equivalence. The most common technique for establishing structural equivalence is factor analysis. Both exploratory and confirmatory factor analysis can be used to address structural equivalence. The former amounts to a comparison of factor loadings (computational details can be found in Van de Vijver and Leung, 1997). Suppose that an instrument to measure organizational commitment is administered to employees in two countries. The same number of factors is extracted in both countries. The solution of one country is then rotated to the solution of the other country. This step is necessary to correct for the rotational freedom in exploratory factor analysis. In the last step of the procedure the agreement is computed for each factor extracted. A common statistic to compute the factorial agreement is known as Tucker's (1951) phi, originally proposed by Burt. This statistic computes the identity of two factors up to a positive multiplying constant. Factors in different countries with identical eigenvalues should have identical factor loadings, whereas factors with different eigenvalues are first corrected by multiplying the loadings with a positive constant so as to equate their eigenvalues. Allowing eigenvalues to differ across cultures before comparing the loadings is based on the reasoning that factors with different reliabilities across cultures can still measure the same underlying construct.

There are two different ways in which factor structures can be compared across cultures. The first procedure involves a pairwise comparison of factor structures across all countries. This strategy can quickly become cumbersome is the number of countries involved is large. A comparison of n countries involves $n \times (n - 1)/2$ comparisons. Comparing ten cultures already amounts to forty-five comparisons. The second procedure involves a comparison of all cultures to a single target culture (in which an instrument measuring the instrument was developed and validated) or to a pooled solution (to which all countries contribute either equally or weighted by their sample size). If the

number of countries is relatively small, a researcher may decide to compare each country to the pooled solution of the other countries to avoid that a country contributes to the overall solution to which it is compared. The number of comparisons to be made is equal to the number of countries involved; a ten-country study would involve ten comparisons. The procedure in which a single solution for all countries is used as reference has become standard both in exploratory and confirmatory factor analysis. The reasons for this choice are computational simplicity and scientific parsimony (a single model accounts for the data in all countries). However, the procedure is problematic if there are homogeneous clusters of countries with different solutions. Suppose that we administer an instrument to measure depression in various countries and that the items cover both somatic and psychological symptoms of depression. It is known from the literature that various (non-western) cultures are less likely to endorse the psychological symptoms than the somatic symptoms (Van de Vijver and Tanaka-Matsumi, 2008). It may well be that the instrument is unidimensional in western cultures and bidimensional in non-western cultures. Pairwise solutions are better equipped to identify such homogeneous clusters. A cluster analysis of factorial agreement indices would show the different clusters which is more difficult to find in the analysis of a pooled solution.

Confirmatory factor analysis follows a different procedure. Compared with the exploratory factor analytic procedure, the testing of structural equivalence using confirmatory factor analysis is based on more rigorous statistical procedures and includes more parameters than factor loadings. Suppose that our scale of organizational commitment measures two correlated factors in both countries. The evaluation of equivalence in a confirmatory factor analysis consists of a number of hierarchically ordered tests. The first step tests whether the factor analytic solutions in the two countries have the same configuration which means that the same indicators should load on the same factors. This constellation is called "configural invariance." Assuming that an acceptable fit is found for this model, we can proceed to the next step by selecting parameters of the model that should be identical across cultures. It is customary to test the identity of factor loadings in the next step ("measurement weights"), followed by a test of the identity of regression intercepts of the observed variables on their latent factors, identity of factor covariances, the identity of the structural residuals (i.e., identity of error components of the latent factors), and finally the identity of measurement residuals (i.e., identity of the error components of items). Examples from the organizational literature can be found in Ployhart *et al.* (2003) and Vandenberg and Lance (2000).

In our view, there are two kinds of problems with the use of structural equation modeling in cross-cultural organizational research. The first issue involves the assumption (often implicitly made in empirical applications of invariance tests) that a positive outcome of a test of invariance demonstrates that there is no bias in the instrument. The assumption is also used in the context of differential item functioning. An instrument from which all bias items have been removed is assumed to show valid cross-cultural score differences. The assumption is not correct. It is correct to argue that a failure to find invariance points to the presence of bias; however, it is quite possible that there is bias even if a test of invariance produces favorable results. The problem is a consequence of the absence of a rigorous test of construct bias in standard tests of invariance. An instrument that measures filial piety according to its western conceptualization leaves out important aspects of the concept in a non-western context, even if the instrument would show the highest level of cross-cultural invariance. There is a second and related assumption in invariance testing that also requires scrutiny; we refer here to the assumption, again often implicit, that a comparison of means based on instruments that have shown invariance shows cross-cultural differences that are only related to the target construct. The assumption is problematic because the influence of sources of method variance with a pervasive influence on items, such as acquiescence or social desirability, may not have been ruled out. It should be pointed out that the problematic nature of these assumptions is not a consequence of the statistical properties of structural equation models but of their current usage. There are indeed examples of cross-cultural studies in which structural equation modeling is used

to examine the influence of acquiescence on cross-cultural score differences (Welkenhuysen-Gybels, Billiet, and Cambré, 2003).

A second problem in the use of structural equation modeling in cross-cultural studies involves the use and interpretation of fit statistics. There is a rich literature on fit statistics. Cheung and Rensvold (2002) conducted a simulation study to evaluate various fit statistics to test invariance in two-country comparisons. They suggest the use of increases in Bentler's comparative fit index, Steiger's gamma hat, and McDonald's noncentrality index in invariance testing. Their results, though very useful, should be complemented by more empirical studies in which the suitability of these guidelines are tested and by more Monte Carlo studies in which extensions to commonly applied fit indices such as the AGFI and to larger numbers of countries are studied. We do not yet know how we can adequately evaluate model fit in cross-cultural projects that involve dozens of countries. It has been proposed that an alternative way of overcoming fit problems could be the use of so-called item parcels (e.g., Little *et al.*, 2002). Items are combined in parcels so as to reduce the impact of item particulars on model fit such as differential skewness and kurtosis of items across countries. Cross-cultural differences in these distributional properties can lead to a poor fit, although they may be minor and psychologically trivial. The use of item parcels could hold an important promise for cross-cultural research. However, their current usage is hampered by two factors. The first is the absence of generally accepted ways as to how items should be clustered. The second is related to the first; it has been demonstrated that bias in items may remain unnoticed if biased items are included in parcels with unbiased items (Meade and Kroustalis, 2006).

Explaining cross-cultural differences

Experienced cross-cultural researchers know that it is often easier to find significant cross-cultural differences in mean scores than to provide a conclusive interpretation of these differences. An important methodological aspect of cross-cultural research is to rule out alternative interpretations (Campbell, 1986). For example, suppose that a study shows that turnover intention is higher among employees in a US company than in a Japanese company. A first interpretation could be that the observed difference reflects a real cross-cultural difference which is in line with the lower labor market mobility of Japanese workers (as compared to American workers). However, various alternative interpretations could be offered. The first one would be that the construct or particular items are biased (e.g., the factor structure of the instrument is not the same in the two countries or some items are inadequate for the American employees). It could also be that the nature of the companies was different (e.g., the Japanese company is known to be a good, well paying employer), that the educational level of the employees was different (e.g., the Japanese employees were less schooled which makes them less mobile), or that the Japanese workers were less inclined to admit that they consider to quit their job. A common way to examine the validity of these interpretations is to include relevant operationalizations in the research so that its impact can be investigated. For example, a social desirability questionnaire is administered and a covariance analysis is carried out to examine whether cross-cultural differences are significant after social desirability differences in the two countries have been taken into account. The validity of our original interpretation of the cross-cultural differences (in terms individualism – collectivism) increases when we can rule out more alternative interpretations.

The search for validations of cross-cultural differences has an interesting and possibly unexpected corollary. Suppose that the differences in the above example are no longer significant if country differences in social desirability have been taken into account. Such a finding has an important psychological implication: the cross-cultural differences in turnover intention have to be seen as differences in social desirability. Japanese and American employees with the same level of social desirability are expected to have the same turnover intention. We may think that we observe cross-cultural differences in turnover intention, but what we actually observe are correlates of cross-cultural differences in social desirability. The cross-cultural literature contains various examples of how cross-cultural differences in target variables are shown to be reflections of other

variables. For example, many differences between immigrant groups and mainstreamers in the acculturation literature are a function of the differences in socioeconomic status or education of the groups. Arends-Tóth and Van de Vijver (2008) found that the more traditional family values of non-western immigrant groups in the Netherlands (as compared to the Dutch mainstream group) can be largely explained by differences in education. Immigrants and mainstreamers with the same educational background do not show substantial differences in family values.

The methodological approach to validate interpretations of observed score differences in cross-cultural studies is known as "unpackaging" (Bond and Van de Vijver, 2008; Whiting, 1976). The idea is that observed score differences in target variables should be the starting point of further inquiry and that an examination of the antecedents of these differences is required; the differences should be "unpackaged." This process of unpackaging may involve the confirmation of intended interpretations (e.g., a measure of individualism – collectivism is administered and can statistically account for the observed cross-cultural differences in turnover intention); the process may also involve the disconfirmation of non-target explanations (e.g., the educational level of employees is measured so that we can statistically examine whether cross-cultural differences in education can explain away the differences in turnover intention). If researchers have a larger number of cultural groups, multilevel analyses can provide a powerful and elegant alternative for addressing bias issues. Conceptually similar to the "unpackaging," culture level variables can be used to examine whether they explain the observed cultural differences at the individual level. Although equivalence and multilevel approaches are often treated as separate topics, both approaches can be used to address questions of bias and equivalence (if large samples are available; see Fontaine, 2008).

Multilevel issues in organizational research

The literature on multilevel issues in organizational research has a comparatively long tradition. This is not surprising, given that managers have to deal with issues at the level of individuals, dyads, work groups, departments, and whole organizations. If organizational theories do only apply at one level (let us say the individual) and are misspecified at another level (work group or department), then organizational survival might be threatened and the manager could potentially lose his/her job if such theories were applied at the wrong level. Interest in multi-level research has increased exponentially over the last few decades with an associated sophistication and diversification of approaches (Kozlowski and Klein, 2000). Special issues on level issues in prestigious journals such as *Academy of Management Review*, *Leadership Quarterly*, and *Journal of International Business Studies* have been published, and there have been dedicated books and book series on the topic from organizational perspectives (e.g., Dansereau Alutto, and Yammarino, 1984; Klein and Kozlowski, 2000a; Yammarino and Dansereau, 2002–2007). The conceptual and statistical models that have been developed allow for an integrated treatment of the three basic issues of multilevel modeling mentioned before (What is the appropriate level of a theory (and data)? Is there a change in meaning of the same construct after (dis)aggregation?) Nevertheless, the research practice shows a more fragmented picture.

Identifying the appropriate level of theory and data

The first step for any research project should be the identification of the appropriate level to which generalizations should be made (Klein, Dansereau, and Hall, 1994). Are we proposing a theory that explains the motivation of individuals, interaction patterns between individuals in teams or the behavior of larger organizations? Although this may seem rather straightforward, the definition of the appropriate level can often be quite ambiguous. For example, many constructs such as justice perceptions, self-efficacy, or affect were thought to capture individual-level constructs, but, more recently, researchers have demonstrated these processes can also be described at higher levels; see work on justice climate (Colquitt, Noe, and Jackson 2002),

group efficacy (Bandura, 1997), and group affect (George and James, 1993).

To help with the development of theory and research, Klein, Dansereau, and Hall (1994) outlined three alternative assumptions underlying any theoretical model: *homogeneity*, *independence*, and *heterogeneity*.

Homogeneity (or wholes in Dansereau, Alutto, and yammarino's (1984) terminology) refers to the homogeneity of subunits within higher level units. Variability within units is seen as error. Using individuals within groups as an example, "group members are sufficiently similar with respect to the construct in question that they may be characterized as a whole" (Klein, Dansereau, and Hall 1994, p. 199). A single value or characteristic is then seen as sufficient to describe the group as a whole. Aggregation of responses by individuals within groups is justified if individuals within a specific unit agree with each other about the psychological meaning of the construct. In the theoretically ideal case, true variation only occurs between groups or units, but not within (James, 1982) and true effects exists only between units, phenomena are shared and identical within units and within-unit variability is error. In cross-cultural psychology, the common definition of culture as a shared meaning system (e.g., Hofstede, 1980, 2001; Rohner, 1984) would follow a homogeneity assumption.

The second assumption is *independence*. Subunits are independent from higher-level units. For example, individuals would be free of group influence. This assumption is made by many statistical tests (e.g., individual scores are independent from each other). This assumption treats group membership as irrelevant and the only true variation is between individuals (e.g., individual differences). Psychological approaches to human behavior have often been criticized for strongly adhering to this assumption (Sampson, 1981).

The final assumption is called *heterogeneity*, "*frog-pond*", *within-group* or *parts effect* (e.g., Dansereau et al., 1984). Comparative or relative effects are theorized and absolute effects are not important. A frog may be comparatively small in a big pond, but the same frog would appear large if the pond was smaller. The main assumption is therefore that effects are context-dependent, with any score depending on the respective level of scores in the unit of interest. The classical example is social comparison processes (Festinger, 1954). Individuals compare themselves with others and the standing relative to the standard or referent is important. Therefore, individuals vary within groups, the group itself is a meaningful entity and necessary as a contextual anchor, but variations between groups are not the key focus.

These theoretical issues have implications for both operationalization of constructs and sampling. Having theoretically defined an intended level of analysis, researchers need to decide how to best operationalize their theoretical constructs. Composition models (Chan, 1998) address how constructs can be measured at various levels. They "specify the functional relationship among phenomena or constructs at different levels of analysis … that reference essentially the same content but that are qualitatively different at different levels" (Chan, 1998, p. 234). These models are helpful for conceptual precision in construct development and measurement since they deal with the content of dimensions and item wording.

Most constructs can be defined and investigated at various levels. Values as an example have been measured at the level of the individual, organization, and nation. At the level of the individual we would deal with an individual construct, whereas at the organization or nation level it reflects a collective construct. This distinction between individual and collective constructs is important (Morgeson and Hofmann, 1999). Individual-level constructs pertain to individuals and may reflect neuro-physiological or genetic processes, individual learning or specific and idiosyncratic life experiences. It may also be possible to describe the average level of any individual-level construct within a particular group. Aggregations of individual level constructs are possible, but the nature and function of such aggregates remains purely at the individual level.

In contrast, collective constructs clearly operate at the higher collective level and can not be broken down to the individual level. Morgeson and Hofmann (1999, p. 253) highlight that: "Collective structures emerge, are transmitted and persist through the actions of members of the collective (or the collective as a whole)." Speaking of the

Table 18.3 A classification of aggregate and collective constructs

Name of model	Level of observation	Agreement within group	Referent
Selected score model	Collection of individuals	Not necessary	Individual
Summary index model	Collection of individuals	Not necessary	Individual
Dispersion model	Collection of individuals	Not necessary	Individual
Referent shift model	Collective	Necessary	Aggregate
Aggregate model	Collective	NA	Aggregate
Consensus model	Fuzzy	Necessary	Individual

"collective climate" of an individual, for example, would be inappropriate and most people would agree that this does not make sense. Collective climate needs a group context to become meaningful. As Morgenson and Hofmann (1999, p. 252) put it:

> Mutual dependence (or interdependence) between individuals creates a context for their interaction. This interaction, in turn, occasions a jointly produced behavior pattern, which lies between the individuals involved. Collective action, thus, has a structure that inheres in the double interact rather than within either of the individuals involved. As interaction occurs within larger groups of individuals, a structure of collective action emerges that transcends the individuals who constitute the collective.

We briefly describe six different composition models. The statistical properties and origins of the model are more fully described in Chen, Mathieu, and Bliese (2004), Fischer (2008) and Hofmann and Jones (2004). We will describe these models in relation to individuals, organizations and nations, although these models are applicable to any other theoretical level (dyads, teams, departments, industries, regions, etc.).

The first three models in table 18.3 describe collections of individuals. The *selected score model* refers to an aggregate defined through a specific score at the individual level. This model most often applies to boundary conditions. For example, in the team productivity literature, team performance might be constrained by the lowest performing individual (Steiner, 1972). Therefore, one selected score would identify the higher level score, but the score is still at the level of the individual.

The *summary index model* describes groups through the aggregate of a variable of interest at the individual level. We could, for example, measure the personality of all group members and then assign the average personality profile to each group. Therefore, the mean of an individual level variable is assigned to a whole work group. According to Hofmann and Jones (2004), the summary index model reflects the mean or sum of a construct for a collection of *individuals*, but it does not provide any meaningful information about the collective (work group in our example). These mean scores are therefore best interpreted as the central tendency of individuals.

The final individual level model is the *dispersion model*. Here, the variability or distribution of characteristics or properties rather than their central indices are of interest. It is similar to the previous summary index model in that it represents descriptive statistics of individuals within a unit or group. This variability is most commonly assessed using indicators of within-group variance (e.g., Naumann and Bennett, 2001). Value diversity within groups can be assessed with dispersion models (Williams and O'Reilly, 1998).

Collective constructs can be measured using the next three models in table 18.3. According to Hofmann and Jones (2004), both the referent-shift models and aggregate properties models provide clear and non-ambiguous assessment of true collective constructs. *Referent-shift models* were developed in climate research (Chan, 1998; Glick, 1985) to avoid conceptual confusions between individual (psychological) and organizational (collective) climate. Referent-shift models ask individuals to answer items focusing on the higher-level unit of investigation (work group or organization). Therefore, the referent is changed from "I" to "we" or "this group." Hence, a value item would look like "In this workgroup, people value power."

An essential step for referent-shift models is the assessment of agreement prior to aggregation. Data should only be aggregated if there is sufficient agreement (see below). Hence, the marked characteristics of this model are (a) focusing responses of individuals on the higher unit (instead of self-reports) and (b) an evaluation of agreement to justify aggregation (since agreement would indicate a collective construct). Referent-shift models are similar to summary-index models in that both require reports of individuals. However, summary-index models measure self-reports of individuals about their own characteristics, attitudes, abilities or values and these reports are aggregated without assessing agreement.

The second model of collective constructs is the *aggregate properties model*. This is the simplest model in that the construct directly reflects the higher unit. For example, the number of individuals working in an organization, the number of hierarchical levels or distributions of experts throughout departments are clear indicators of organizational-level characteristics. Expert ratings are also valid (e.g., ratings on organizational performance or innovation characteristics by the CEO).

The final model in this typology is the *consensus model*. Compared to the other two models, it is conceptually more complex, ambiguous or fuzzy (Hofmann and Jones, 2004). It may indicate a collective construct, since it is essentially an individual-level construct, but for which agreement exists. For example, if ratings of an item such as "I am happy" were found to be homogeneous within work groups or organizations, it would be justified to aggregate the scores to a higher level (this dependency at the individual level would also lead to biases and wrong statistical estimates at the individual level if not aggregated; Barcikowski, 1981; Bliese and Hanges, 2004; Kenny and Judd, 1986). Therefore, this model is similar to both the summary index model (by using individual-referenced items) and referent-shift consensus model (by showing sufficient agreement).

Hofmann and Jones (2004) prefer referent-shift models over direct-consensus models because direct-consensus models are ambiguous by providing an index of the shared level of individual-level characteristics within the culture, whereas the referent-shift consensus model represents the collective construct directly. Hofmann and Jones (2004) treat direct-consensus models as (indirect) markers for true collective constructs with referent-shift models being preferred for measuring collective constructs (Klein, Dansereau, and Hall, 1994; Kozlowski and Klein, 2000; Morgeson and Hofmann, 1999).

Assessment of agreement

Agreement is essential for developing true collective construct measures. A number of indicators are available and there has been a healthy debate in the literature about the appropriateness and empirical cut-off criteria for sufficient agreement that justify aggregation. One of the older and widely used indices is r_{wg}, developed by James, Demaree and Wolf (1984, 1993). This index focuses on consensus or agreement within a single unit; for example, a work group. This index compares the variability of a variable within a work group to some expected variability. If the observed variability is substantially smaller than the expected variance, the resulting value of r_{wg} is closer to 1, suggesting high agreement and that aggregation is possible. The index ranges from 0 to 1, although negative or values larger than 1 are possible (James, Demaree, and Wolf 1984; Klein and Kozlowski, 2000b). In contrast to reliability estimates that are based on the inter-item correlation, this index uses information about the variability (variance) within units.

Over the years, this index has been used widely but also has been strongly criticized. Brown and Hauenstein (2005) discussed a number of shortcomings of this indicator, among others the dependence on the number of scale options (the more scale options, the higher the agreement with everything else being equal), the dependence on the sample size (the greater the sample size, the higher the agreement, everything else being equal) and problems with the assumption of the null distribution (which is typically a rectangular distribution). They proposed an alternative measure a_{wg}. The maximum possible variance at the mean is being used as the null distribution. Agreement is then calculated as the 1 minus twice the observed variance divided by the maximum possible variance. The range of the index varies between −1 and 1.

A value of 1 means perfect agreement, a value of −1 indicates perfect disagreement and a value of 0 indicates that the variability is fifty percent of the possible variance at the mean. There are no statistical significance tests associated with a_{wg}. A .70 cut-off value has been proposed as a heuristic for moderate agreement, with values of less than .59 being seen as unacceptable if the construct is supposed to reflect group-level constructs (Brown and Hauenstein, 2005). Previous research has focused on agreement around specific and well-defined aspects in small groups within organizations. The critical values calculated by Brown and Hauenstein (2005) are based only on groups smaller than twenty; consequently those guidelines might be overly conservative with larger groups (such as organizations or nations). However, the index is a significant improvement since it overcomes several shortcomings of the widely used r_{wg}.

A second class of statistics to evaluate the extent to which perceptions are shared are intra-class correlations (ICC) (James, 1982; Shrout and Fleiss, 1979). Two types are commonly in use, ICC(1) and ICC(2). The first is essentially based on a random one-way analysis of variance and provides an estimate of the proportion of the total variance of a measure that is explained by unit membership (Bliese, 2000). A second interpretation of ICC(1) is as an estimate of the extent to which any one rater may represent all the raters within a group, the question of whether raters are interchangeable (James, 1982). The advantage of ICC(1) over other estimates such as eta-squared is that it is independent of group size (Bliese, 2000; Klein and Kozlowski, 2000b).

ICC(2) is used to answer the question about reliability of group means within a sample. ICC(2) values like any measure of reliability should exceed .70 to be judged as acceptable. This index is a variant of ICC(1), basically ICC(1) adjusted for group size (Bliese, 2000). Similar to other measures of reliability (e.g., Cronbach's alpha), the larger the group size, the larger ICC(2). This is based on the logic that group means based on many people per group are more stable and reliable than group means derived from only a few members. One important difference between r_{wg} and ICC is that r_{wg} focuses on agreement within each group separately (yielding one estimate for each group separately), whereas ICC compares the variability within groups to the variability between groups (yielding one estimate across all groups). One problem that may emerge is that the interrater agreement varies substantially between groups. This can be incorporated in theoretical models as the concept of climate strength (Schneider, Salvaggio, and Subirats, 2002) and its effects can be tested (Colquitt, Noe, and Jackson 2002; Lindell and Brandt, 2000).

The identification of the appropriate level of data and analysis also has implications for sampling. Theoretical concerns are important again. Many nations have long histories of immigration and cultural heterogeneity (US, Canada, India, Switzerland, Malaysia, etc.), whereas other nations have been traditionally been more homogeneous in their cultural make-up (Japan, France, Portugal, etc.). Economic migrants also increase cultural diversity in many nations around the world. Rohner (1984) argued that cultural systems consist of equivalent and complimentary meaning systems. Researchers therefore need to identify those elements that are equivalent (shared by all cultural insiders) and those that are equivalent (where cultural knowledge is specific to certain roles and groups). Researchers should sample their research participants in line with the focus of their study. In the case of multicultural samples due to presence of minorities, migrants or the organizational context (multinationals, subsidiaries), indices of dispersion can be included in the theoretical model (e.g., Fischer *et al.*, 2005). In these situations it can be tested whether cultural effects are stronger if they are widely shared within a nation. The above-mentioned indicators of agreement can be used and implemented in research design and analysis. It is also possible to develop models of cultural dispersion to explain cultural phenomena. Gelfand, Nishii, and Raver (2006) developed a multilevel theory of tightness-looseness to account for variability in individual and organizational variables. These theoretical innovations are exciting avenues to explain cultural phenomena as well as addressing issues of increasing cultural change.

A variance approach to levels research

Dansereau, Alutto, and Yammerino, (1984) developed a variance-based approach to test the

appropriate level of a theory. Their "within and between analysis" (WABA) is a complex set of statistical techniques based on ANOVA logic to represent relationships. WABA can be used to test (a) the extent to which a construct varies within- or between-units (WABA I) and (b) to which extent two or more variables covary primarily within-units, between-units or both within- and between-units (WABA II). Therefore, WABA I can be used to assess to what extent variables measured at a lower level can be aggregated to a higher level. WABA II then offers a set of techniques to analyse the appropriate level of the relationships among variables. Data for each variable are divided into within-entities (deviation from the unit average) and between-entity (between unit averages). There are three basic steps. First, each variable is examined to what extent it varies mainly between groups (suggesting homogeneity within groups), within groups (suggesting heterogeneity within groups) or both between and within groups (suggesting individual differences rather than homogeneity or heterogeneity). Second, the relationships between variables are examined to see whether correlations are mainly a function of between-group covariances, within-group covariances and within- and between covariances. These two steps are then assessed for consistency and integrated to draw some overall conclusions about the most appropriate level of analysis (see Dansereau, Alutto, and Yammarino, 1984; Yammarino and Markham, 1992).

The unique aspect of WABA is the availability of tests of practical significance (*E*, *A*, and *R* tests) in addition to statistical significance (*t*, *F* and *Z* tests). These tests of practical significance are geometrically based and do not rely on sample size (degrees of freedom). WABA can also be used to study moderator effects (termed multiple relationship analysis MRA) (Schriesheim, Castro, and Yammarino, 2000).

WABA is a fairly flexible technique that has relatively few assumptions (essentially all the assumptions of ANOVA and regression analyses; see Castro, 2002). The technique does not make any assumptions about the appropriate level of relationships and researchers can test alternative levels of analysis. Therefore, dependent and independent variables are not constrained to any particular level of analysis and researchers can explore the most appropriate level. This is also a limitation since the analyses are completely data driven and testing all possible relationships may not make much theoretical sense (George and James, 1993). However, for the final test of bivariate relationships (WABA II), the relationships need to be at the same level. MRA also requires that the moderator is at a higher level (see Castro, 2002). The practical tests (the E-test in WABA I) has been criticized for being too conservative when group sizes increase (Bliese and Halverson, 1998). With large groups (e.g., using organizations or nations), achieving practical significance becomes difficult. George and James (1993) also noted that restrictions of between-group variance (e.g., when sampling multiple teams from one organizations) may lead to misspecifications of the WABA I equations. A final limitations that might be of particular interest for cross-cultural researchers is that WABA is not applicable in cases in which the relationship between variables *x* and *y* differs depending on the group (a person x situation/group interaction). If the relationships differ significantly across groups, the fundamental WABA equation will be meaningless since it follows the logic of ANCOVA that assumes equality of regressions lines (George and James, 1993). This is a concern for cross-cultural researchers, since it is a well-known phenomenon that relationships can be culture-specific (see, for example, the discussion of functional and structural equivalence above). Nevertheless, WABA has much to be recommended for cross-cultural research, since the technique can integrate various seemingly divergent multi-level perspectives (Dansereau and Yammarino, 2006).

Assessing changes of meaning of the same construct after aggregation

The previous section was concerned with the determination of the appropriate level of analysis. The implicit assumption was that the meaning of constructs remains the same. Organizational researchers using the methods described above have been less concerned with meaning changes during aggregation. In contrast, this has been a central concern for the approaches that are discussed next.

It is important to note that the methods described in these two sections have been independently developed and an integration is needed (see Peterson and Castro, 2006). Methods that were discussed in the bias and equivalence section can be used to address changes in meaning since it is a different form of equivalence (equivalence of meaning across levels).

Establishing factor structures at more than one level

Hofstede (1980) using a large cross-cultural dataset showed that the factor structures at the individual and national level can be different. This finding has led to a substantive interest among cross-cultural researchers in the structure of constructs at various levels. As discussed previously, WABA shows that within- and between structures are independent and can lead to completely different relationships. There are three statistical techniques that have been used for establishing equivalence across levels: multidimensional scaling, exploratory factor analysis and confirmatory factor analysis (for a more detailed description see Fontaine, 2008; Fischer, 2008; van de Vijver and Leung, 1997).

First, it would be important to analyze the structure at the individual level. As discussed before, within and between-group covariances are mathematically independent. Therefore, it is best to compare factor structures pairwise between nations or better, compare each nation with a pooled factor structure that gives equal weight to each group (and removes the between-group covariance component). Using the total covariance matrix across all participants irrespective of groups will lead to a mixing of within and between-effects. This should be avoided since it blurs the relative structures. Once an acceptable factorial structure (using acceptable agreement across individual solutions, see above) is found across cultural groups, it can be tested to what extent this individual-level structure has a comparable structure at the aggregate level. The aggregated between group correlation or covariance structure is factor analyzed or analyzed using multi-dimensional scaling. This between-group structure is then compared to the average individual-level solution (Muthén, 1994; van de Vijver and Poortinga, 2002). As we have discussed previously, it would important to test within-group agreement and between-group variability prior to aggregation. Sufficient between-group variability is obviously necessary, otherwise there would be nothing to model at the higher level. Therefore, this step of assessing between-group variability (recommending mostly ICC(1)) is included in most recommendations of multi-level factor analysis (e.g., Muthén, 1991, 1994; van de Vijver and Poortinga, 2002).

The comparison of individual solutions at the individual level followed by a comparison with the aggregated matrix is a necessary step for all three techniques (although programmes like MPlus now allow simultaneous estimation of within- and between-group structures, see Fontaine, 2008). For MDS and EFA, an additional step is necessary. As discussed above, the structures need to be rotated to maximal similarity to allow comparisons across levels. Factorial agreement indices are available (see van de Vijver and Leung, 1997) and allow estimation of the similarity at the factor-level. CFA does not require this rotational sub-step. CFA is also more sophisticated, in that it allows for theory-driven constraints of parameters across levels and provides statistical tests for differences of individual parameters across levels. However, a drawback of CFA is that this technique has more assumptions (e.g., multivariate normality), fit indices are sample size dependent and there is a continuing debate about appropriate indicators of fit (see the discussion above).

In summary, the question of changes in meaning of constructs across levels due to aggregation is contentious, but can readily be addressed through multidimensional scaling or factor analysis at both levels. The structures can then be compared and inferences about the similarity or differences can be made. The previous section on the appropriate level of analysis has also demonstrated that it is theoretically possible that structures will be different since the within- and between-group covariance matrices are mathematically independent. These MDS and factor analytical techniques can be implemented without examining agreement or variance components. However, the two questions are complementary and ideally should be integrated.

Relationship between different constructs across levels

The final question addresses how different constructs are related across levels. We can distinguish three broad types of models: single-level models, cross-level models and homologous multi-level models (Klein and Kozlowski, 2000b). Single-level models are the most common models in that they are dealing with relationships between construct at one level of theory only. This level may be the individual, group, organization or nation-level. Psychologists and management researchers are most familiar with individual-level models, management researchers often deal with models at the team or organizational level and cross-cultural psychologists and sociologists are also familiar with models at the nation-level. Since single level models do not deal with constructs at a higher or level of analysis, they are straightforward analyzable using traditional analytical techniques such as correlation, multiple regression or structural equation modeling. If single-level models are conceptualized at a higher level and based on aggregation of lower-level data, all the steps addressed in relation to the first two questions need to be followed.

Cross-level models are the most complex models since they conceptualize relationships between variables across different levels. Organizational researchers are most familiar with top-down approaches that model effects of higher level variables on lower level variables (e.g., organizational climate influencing employee job satisfaction or performance). The alternative process of emergent or bottom-up processes is equally plausible, but empirical research on such processes is as yet sparse (Kozlowski and Klein, 2000). This is an area which has much potential for further theoretical development, particularly since it addresses essential questions such as how collectives develop and can be changed. Such research needs to be time-sensitive since emergent processes are slower and show delayed effects compared to top-down models (Klein and Kozlowski, 2000b). The statistical technique most suited to address emergent processes at this stage is WABA.

However, in the following we will focus on the three broad types of top-down models (Klein and Kozlowski, 2000b, Klein, Dansereau, and Hall 1994): direct effects, moderator and frog-pond cross-level models. The first model conceptualizes and examines direct or main effects of a higher level variable on one variable at a lower level. For example, we could estimate whether macro-economic development or thermal climate at the nation-level affects the willingness of individuals to volunteer within nations (van de Vliert, Huang, and Levine, 2004). In this case, both macro-economic and thermal climate are clear nation-level variables and their effect on the means within nations are estimated. Cross-level direct effects models can be used for unpackaging cultural effects and to investigate bias issues. When a large number of cultural samples is available (ideally twenty or more samples), researchers can first estimate the cross-cultural differences (e.g., using ICC(1)). As discussed above, these differences are often ambiguous to interpret and can arise due to substantive processes as well as a number of biases. If this variability can be explained using variables at a higher level (e.g., individualism-collectivism, national wealth), biases can be eliminated as alternative explanations or the relative effect of potential bias can be estimated (by examining how much variance is unexplained after accounting for the explanatory variables of interest).

Cross-level moderator models are complementary to direct effect models since they additionally examine whether a higher level variable changes the relationship between two lower level variables. More complex models are also possible. For example, Huang, van de Vliert, and van der Vegt (2006) studied whether power distance at the nation-level changed the relationship between employment involvement and participative climate on employee voice (a proactive tendency to make suggestions about improvements) at the organizational level. Therefore, the dependent variable at the organizational level was employee voice, the two independent variables at the organizational level were participative climate and formalized employee involvement, the independent variable at the nation-level was power distance. They found a three-way interaction across levels. Power distance changed the relationship between employee involvement and employee voice, but only if participative climate

is high. In high power distant nations, formalized employee involvement is associated with increased employee voice, but only if there is a strong participative climate. Cross-level moderator models can also be used to address bias issues. For example, acquiescence and extreme responding are forms of method bias that threatens measurement unit and full-score equivalence. Smith and Fischer (2008) tested whether individual differences and culture-level variables together explain variability in these response styles. They found significant interactions, highlighting that individual dispositions and cultural variables have interactive effects on the willingness of respondents to acquiesce or express extreme opinions in survey research. For example, interdependent individuals in contexts in which it is acceptable to express affect freely (high affective autonomy) were more likely to agree to items irrespective of content. In contexts that were low on affective autonomy, the level of agreement was low irrespective of the interdependence of individuals.

The final set of cross-level models is cross-level frog-pond models. These models are related to the heterogeneity assumption described above since it models the effects of individual group members standing within a group on individual-level outcomes. An example is the relationship between performance of individuals and their self-efficacy, depending on the average level of performance within the team. In a high performing team, an individual with less than average work performance is likely to experience lower esteem. However, if the same individual was placed in a low performing team, his/her previously mediocre performance would be above average and the level of self-esteem might improve. In essence, the relative standing within the team is of importance rather than absolute levels. WABA II is best suited to test such frog-pond models (Klein and Kozlowski, 2000b).

The last group of multilevel models discussed here are homologous multilevel models. These models are somewhat similar to single-level models since they do not specify relationships across levels, but only relationships within levels. However, these models also specify that relationships between variables hold at multiple levels of analysis. The great appeal and value of such models for researchers is that generalizations across level can be made, substantially enhancing the generality and applicability of theory. A drawback of these models is that the demand for similar structures and functions across levels leads to abstract and simplified theoretical models that are no longer of any practical value (Klein, Cannella, and Tosi, 1999; Klein and Kozlowski, 2000b). To date, no such model has been proposed and empirically tested. Consequently, these models have much theoretical appeal, but their practical utility and usefulness is yet unproven. Chen, Bliese, and Mathieu (2005), as well as Zyphur and Preacher (in review) have recently proposed conceptual frameworks and statistical procedures for such models and this may help to generate more theory and empirical tests (for a critique of these approaches see Dansereau and Yammarino, 2006).

Conclusion

A sound methodology can enhance the validity of findings. This truism is also true in cross-cultural organizational research. The appropriate uses of methodological tools can help to improve the interpretability of cross-cultural studies. Thus, various sophisticated tools are available to address the question of whether in instrument measures the same in different cultures. Examples are exploratory and confirmatory factor analyses and the numerous techniques that can be employed to identify differential item functioning. We have seen tremendous developments in cross-cultural research methods in the last decades. However, these techniques are not always fully exploited. We still come across too many studies in which cross-cultural differences in means scores are taken at face value without any concern for the comparability of scores across cultures. Progress in cross-cultural organizational research will depend on a combination and integration of sophisticated theorizing and adequate use of the tools that are available. It is remarkable that some methodological considerations have been widely accepted, such as problems with low internal consistencies and interrater reliabilities, while other recommendations regarding the testing of equivalence are often more preached than practiced.

We have paid much attention in the chapter to current developments in multilevel models. We consider these models to be possible spearheads of new developments in cross-cultural organizational behavior research. Multilevel models combine innovations in theory and development. We consider these models to be particularly important because they enable the study of individuals, organizations, and cultures in a joint model. As a consequence, we can now model the interaction of variables at different levels.

We expect that a further integration of theory and methods and a more refined use of methodological tools in cross-cultural research will help to increase the replicability of cross-cultural research findings, to bolster our conclusions against alternative interpretations, and to generate theories that better stand testing in a cross-cultural framework.

References

Aquilino, W. S. 1994. "Interviewer mode effects in surveys of drug and alcohol use", *Public Opinion Quarterly* 58: 210–40.

Arends-Tóth, J. V. and Van de Vijver, F. J. R. 2008. "Cultural differences in family, marital, and gender-role values among immigrants and majority members in the Netherlands", *International Journal of Psychology* (in press).

Au, K. Y. 2000. "Intra-cultural variation as another construct of international management: a study based on secondary data of 42 countries", *Journal of International Management* 6: 217–38.

Barcikowski, R. S. 1981. "Statistical power with group mean as the unit of analysis", *Journal of Educational Statistics* 6: 267–85.

Bandura, A. 1997. *Self-Efficacy: The Exercise of Control*. New York, NY: Freeman.

Barrett, P. T., Petrides, K. V., Eysenck, S. B. G., and Eysenck, H. J. 1998. "The Eysenck Personality Questionnaire: an examination of the factorial similarity of P, E, N, and L across 34 countries", *Personality and Individual Differences* 25: 805–19.

Besser, T. L. 1993. "The commitment of Japanese workers and U.S. workers: a reassessment of the literature", *American Sociological Review* 58: 873–81.

Bhagat, R. S. and McQuaid, S. J. 1982. "Role of subjective culture in organizations: a review and directions for future research", *Journal of Applied Psychology* 67: 653–85.

Bliese, P. D. 2000. "Within-group agreement, non-independence and reliability: implications for data aggregation and analyses", in J. K. Klein and S. W. J. Kozlowski (eds.), *Multilevel Theory, Research and Methods in Organizations. Foundations, Extensions, and New Directions*. San Francisco, CA: Jossey-Bass. pp. 349–381

and Halverson, R. H. 1998. "Group size and measures of group-level properties: an examination of eta-squared and ICC values", *Journal of Management* 24: 157–72.

and Hanges, P. J. 2004. "Being both too liberal and too conservative: the perils of treating grouped data as though they were independent", *Organizational Research Methods* 7: 400–17.

Bond, M. H. and van de Vijver, F. J. R. 2008. "Making scientific sense of cultural differences in psychological outcomes: unpackaging the magnum mysterium", in F. J. R. van de Vijver and D. Matsumoto (eds.), *Research Methods in Cross-Cultural Psychology*. New York: Oxford University Press.

Brown, R. D. and Hauenstein, N. A. 2005. "Interrater agreement reconsidered: An alternative to the r_{wg} indices", *Organizational Research Methods* 8: 165–84.

Camilli, G. and Shepard, L. A. 1994. *Methods for Identifying Biased Test Items*. Thousand Oaks, CA: Sage.

Campbell, D. T. 1986. "Science's social system of validity-enhancing collective believe change and the problems of the social sciences", in D. W. Fiske and R. A. Shweder (eds.), *Metatheory in Social Science*. Chicago, IL: University of Chicago Press, pp. 108–135.

Castro, S. L. 2002. "Data analytical methods for the analysis of multilevel questions: a comparison of intraclass correlation coefficients, $r_{wg(j)}$, hierarchical linear modelling, within and between-analysis and random group resampling", *Leadership Quarterly* 13: 69–93.

Chan, D. 1998. "Functional relations among constructs in the same content domain at different levels of analysis: a typology of compositional models", *Journal of Applied Psychology* 83: 234–46.

Chen, G., Bliese, P. D., and Mathieu, J. E. 2005. "Conceptual framework and statistical procedures for delineating and testing multilevel theories of homology", *Organizational Research Methods* 8: 375–409.

Mathieu, J. E., and Bliese, P. D. 2004. "A framework for conducting multilevel construct validation", in F. J. Yammarino and F. Dansereau (eds.), *Research in Multilevel Issues: Multilevel Issues in Organizational Behavior and Processes* 3: 273–303. Oxford: Elsevier.

Cheung, F. M., Leung, K., Fan, R. M., Song, W. Z., Zhang, J. X., and Zhang, J. P. 1996. "Development of the Chinese Personality Assessment Inventory", *Journal of Cross-Cultural Psychology* 27: 118–99.

Cheung, G. W. and Rensvold, R. B. 2002. "Evaluating goodness-of-fit indexes for testing measurement invariance", *Structural Equation Modeling* 9: 233–55.

Cohen, A. 2003. *Multiple Commitments in the Workplace: An Integrative Approach*. Mahwah, NJ: Lawrence Erlbaum.

Cole, E. R. 1979. *Work, Mobility and Participation: A Comparative Study of American and Japanese Industry*, Los Angeles, CA: University of California Press.

Colquitt, J. A., Noe, R. A., and Jackson, C. L. 2002. "Justice in teams: antecedents and consequences of procedural justice climate", *Personnel Psychology* 55: 83–109.

Dansereau, F., Alutto, J. A., and Yammarino, F. J. 1984. *Theory-Testing in Organizational Behavior: The 'Variant' Approach*. Englewood Cliffs, NJ: Prentice Hall.

and Yammarino, F. J. (eds.) 1998. *Leadership: The Multiple-Level Approaches*. Stamford, CT: JAI Press.

and Yammarino, F. J. 2006. "Is more discussion about levels of analysis really necessary? When is such discussion sufficient?" *Leadership Quarterly* 17: 537–52.

Embretson, S. E. 1983. "Construct validity: construct representation versus nomothetic span", *Psychological Bulletin* 93: 179–97.

Eysenck, H. J. and Eysenck, S. B. G. 1975. *Manual of the Eysenck Personality Questionnaire*. London: Hodder and Stoughton.

Festinger, L. 1954. "A theory of social comparison processes", *Human Relations* 7: 117–40.

Fischer, R. 2004. "Standardization to account for cross-cultural response bias: a classification of score adjustment procedures and review of research in JCCP", *Journal of Cross-Cultural Psychology* 35: 263–82.

2008. "Multilevel approaches in organizational settings: opportunities, challenges and implications for cross-cultural research". in F. J. R. Van de Vijver, D. A. van Hemert, and Y. H. Poortinga (eds.), *Individuals and Cultures in Multilevel Analysis*. Mahwah, NJ: Erlbaum.

Ferreira, M. C., Assmar, E. M. L., Redford, P., and Harb, C. 2005. "Organizational behaviour across cultures: theoretical and methodological issues for developing multi-level frameworks involving culture", *International Journal for Cross-Cultural Management* 5: 27–48.

and Mansell, A. 2008. *Levels of Organizational Commitment Across Cultures: A Meta-Analysis*. Paper submitted for publication.

Fontaine, J. R. 2008. "Traditional and multilevel approaches in cross-cultural research: an integration of methodological frameworks" in F. J. R. Van de Vijver, D. A. Van Hemert, and Y. H. Poortinga (eds.), *Individuals and Cultures in Multilevel Analysis*. Mawhaw, NJ: Erlbaum.

Formann, A. K. and Piswanger, K. 1979. *Wiener Matrizen-Test. Ein Rasch-skalierter sprachfreier Intelligenztest* [The Viennese Matrices Test. A Rasch-calibrated non-verbal intelligence test]. Weinheim, Germany: Beltz Test.

Gass, S. M. and Varonis, E. M. 1991. "Miscommunication in nonnative speaker discourse", in N. Coupland, H. Giles, and J. M. Wiemann (eds.), *Miscommunication and Problematic Talk.* Newbury Park, CA: Sage, pp. 121–45.

Gelfand, M. J., Nishii, L. H., and Raver, J. L. 2006. "On the nature and importance of cultural tightness-looseness", *Journal of Applied Psychology* 91: 1225–44.

George, J. M. and James, L. R. 1993. "Personality, affect and behavior in groups revisited: comment on aggregation, levels of analysis and a recent application of within and between analysis", *Journal of Applied Psychology* 78: 798–804.

Glick, W. H. 1985. "Conceptualising and measuring organizational and psychological climate: pitfalls in multilevel research", *Academy of Management Review* 10: 601–16.

Grill, J. J. and Bartel, N. R. 1977. "Language bias in tests: ITPA grammatic closure", *Journal of Learning Disabilities* 10: 229–35.

Hambleton, R. K. 1994. "Guidelines for adapting educational and psychological tests: a progress report", *European Journal of Psychological Assessment* 10: 229–44.

Ho, D. Y. F. 1996. "Filial piety and its psychological consequences", in M. H. Bond (ed.), *Handbook of Chinese Psychology*. Hong Kong: Oxford University Press, pp. 155–65.

Hofmann, D. A. and Jones, L. M. 2004. "Some foundational and guiding questions for multi-level construct validation" In F. J. Yammarino and F. Dansereau (eds.), *Multilevel Issues in Organizational Behaviour and Processes*, vol. 3 Amsterdam: Elsevier, pp. 305–16.

Hofstede, G. 1980. *Culture's Consequences: International Differences in Work-Related Values*. Beverly Hills, CA: Sage.

2001. *Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations*, 2nd edn. Thousand Oaks, CA: Sage.

House, R. J., Hanges, P. J., Javidan, M., Dorfman, P., and Gupta, V. 2003. *GLOBE, Cultures, Leadership, and Organizations: GLOBE Study of 62 Societies*. Newbury Park, CA: Sage.

Huang, X., van de Vliert, E., and Van der Vegt, G. 2006. "Breaking the silence culture: stimulation of participation and employee opinion withholding cross-nationally", *Group and Organization Review* 1: 459–82.

Hui, C. H. and Triandis, H. C. 1989. "Effects of culture and response format on extreme response style", *Journal of Cross-Cultural Psychology* 20: 296–309.

James, L. R. 1982. "Aggregation bias in estimates of perceptual agreement", *Journal of Applied Psychology* 67: 219–29.

Demaree, R. G., and Wolf, G. 1984. "Estimating within-group interrater reliability with and without response bias", *Journal of Applied Psychology* 69: 85–98.

Demaree, R. G., and Wolf, G. 1993. "r_{wg}: An assessment of within-group interrater agreement", *Journal of Applied Psychology* 78: 306–9.

Jensen, A. R. 1980. *Bias in Mental Testing*. New York, NY: Free Press.

Johnson, T. P. 1998. "Approaches to equivalence in cross-cultural and cross-national survey research", *ZUMA Nachrichten Spezial* 3: 1–40.

Kenny, D. A. and Judd, C. M. 1986. "Consequences of violating the independence assumption in analysis of variance", *Psychological Bulletin* 99: 422–31.

Klein, K. J., Cannella, A., and Tosi, H. 1999. "Multilevel theory building: Benefits, barriers and new developments", *Academy of Management Review* 24: 243–8.

Dansereau, F., and Hall, R. J. 1994. "Level issues in theory development, data collection and analysis", *Academy of Management Review* 19: 195–229.

and Kozlowski, S. W. J. (eds.). 2000a *Multilevel Theory, Research and Methods in Organizations. Foundations, Extensions, and New Directions*. San Francisco, CA: Jossey-Bass.

and Kozlowski, S. W. J. 2000b. "From micro to macro: critical steps in conceptualizing and conducting multilevel research", *Organizational Research Methods* 3: 211–36.

Kozlowski, S. W. J. and Klein, J. K. 2000. "A multilevel approach to theory and research in organizations: contextual, temporal and emergent processes", in J. K. Klein and S. W. J. Kozlowski (eds.), *Multilevel Theory, Research and Methods in Organizations. Foundations, Extensions, and New Directions*. San Francisco, CA: Jossey-Bass, pp. 3–90.

Lincoln, J. R. and Kalleberg, A. L. 1990. *Culture, Control and Commitment: A Study of Work Organization and Work Attitudes in the United States and Japan*. New York, NY: Cambridge University Press.

Lindell, M. K. and Brandt, C. J. 2000. "Climate quality and climate consensus as mediators of the relationship between organizational antecedents and outcomes", *Journal of Applied Psychology* 85: 331–48.

Little, T. D., Cunningham, W. A., Shahar, G., and Widaman, K. F. 2002. "To parcel or not to parcel: exploring the question, weighing the merits", *Structural Equation Modeling: A Multidisciplinary Journal* 9: 151–73.

Malpass, R. S. 1977. "Theory and method in cross-cultural psychology", *American Psychologist* 32: 1069–79.

McCrae, R. R. and Costa, P. T. Jr. 1997. "Personality trait structure as a human universal", *American Psychologist* 52: 509–16.

Meade, A. W. and Kroustalis, C. M. 2006. "Problems with item parceling for confirmatory factor analytic tests of measurement invariance", *Organizational Research Methods*, 9: 369–403.

Meyer, J. P. and Allen, N. J. 1991. "A three-component conceptualization of organizational commitment", *Human Resource Management Review* 1: 61–89.

Stanley, D. J., Herscovitch, L. and Topolytsky, L. 2002. "Affective, continuance, normative commitment to the organization: a meta-analysis of antecedents, correlates and consequences", *Journal of Vocational Behavior* 61: 20–52.

Morgeson, F. P. and Hofmann, D. A. 1999. "The structure and function of collective constructs: implications for multilevel research and theory development", *Academy of Management Review* 24: 249–65.

Muthén, B. O. 1991. "Multilevel factor analysis of class and student achievement components", *Journal of Educational Measurement* 28: 338–54.

1994. "Multilevel covariance structure analysis", *Sociological Methods and Research* 22: 376–98.

Naumann, S. E. and Bennett, N. 2001. "A case for procedural justice climate: development and test of a multilevel model", *Academy of Management Journal* 43: 861–89.

Oakland, T., Gulek, C., and Glutting, J. 1996. "Children's test-taking behaviors: a review of literature, case study, and research of children", *European Journal of Psychological Assessment* 12: 240–46.

Peterson, M. F. and Castro, S. L. 2006. "Measurement metrics of aggregate levels of analysis: implications for organization culture research and the GLOBE project", *Leadership Quarterly* 17: 506–21.

Piswanger, K. 1975. *Interkulturelle Vergleiche mit dem Matrizentest von Formann* [Cross-cultural comparisons with Formann's Matrices Test]. Unpublished doctoral dissertation, University of Vienna, Vienna.

Ployhart, R. E., Wiechmann, D., Schmitt, N., Sacco, J. M., and Rogg, K. 2003. "The cross-cultural equivalence of job performance ratings", *Human Performance* 16: 46–79.

Poortinga, Y. H. 1989. "Equivalence in cross-cultural data: an overview of basic issues", *International Journal of Psychology* 24: 737–56.

and Van de Vijver, F. J. R. 1987. "Explaining cross-cultural differences: bias analysis and beyond", *Journal of Cross-Cultural Psychology* 18: 259–82.

Raudenbush, S. W. and Bryk, A. S. 2002. *Hierarchical Linear Models*, 2nd edn. London: Sage.

Rohner, R. P. 1984. "Toward a conception of culture for cross-cultural psychology", *Journal of Cross-Cultural Psychology* 15: 111–38.

Sampson, E. E. 1981. "Cognitive psychology as ideology", *American Psychologist* 7: 730–43.

Schneider, B., Salvaggio, A. N., and Subirats, M. 2002. "Climate strength: a new direction for climate research", *Journal of Applied Psychology* 87: 220–9.

Schriesheim, C. A., Castro, S. L., and Yammarino, F. J. 2000. "Investigating contingencies: an examination of the impact of span of supervision and upward controllingness on Leader-Member Exchange using traditional multivariate within and between entities analysis", *Journal of Applied Psychology* 85: 659–77.

Segall, M. H., Dasen, P. R., Berry, J. W., and Poortinga, Y. H. 1990. *Human Behavior in Global Perspective. An Introduction to Cross-Cultural Psychology*. New York, NY: Pergamon Press.

Shrout, P. E. and Fleiss, J. L. 1979. "Intraclass correlations: uses in assessing rater reliability", *Psychological Bulletin* 86: 420–28.

Singer, E. and Presser, S. 1989. "The interviewer", in E. Singer and S. Presser (eds.), *Survey Research Methods*. Chicago: University of Chicago Press, pp. 245–246.

Smith, P. B., Bond, M. H., and Kagitcibasi, C. 2006. *Understanding Social Psychology Across Cultures: Living and Working in a Changing World*. Thousand Oaks, CA: Sage.

and Fischer, R. 2008. "Acquiescence, extreme response bias and levels of cross-cultural analysis", in F. J. R. van de Vijver, D. A. Van Hemert, and Y. H. Poortinga (eds.), *Individuals and Cultures in Multilevel Analysis*. Mahwah, NJ: Erlbaum.

Fischer, R., and Sale, N. 2001. "Cross-cultural industrial/organizational psychology", in C. L. Cooper and I. T. Robertson (eds.), *International Review of Industrial and Organizational Psychology* vol. 16, pp. 147–94. New York, NY: Wiley.

Steiner, I. 1972. *Group Processes and Productivity*. New York: Academic Press.

Tanzer, N. K., Gittler, G., and Ellis, B. B. 1995. "Cross-cultural validation of item complexity in a LLTM-calibrated spatial ability test", *European Journal of Psychological Assessment* 11: 170–83.

and Sim, C. Q. E. 1991. *Test Anxiety in Primary School Students: An Empirical Study in Singapore*. Research Report 1991/6. Graz, Austria: Department of Psychology, University of Graz.

Tucker, L. R. 1951. *A Method for Synthesis of Factor Analysis Studies*. Personnel Research Section Report No. 984. Washington, DC: Department of the Army.

van de Vijver, F. J. R. 2003. "Test adaptation/translation methods", in R. Fernández-Ballesteros (ed.), *Encyclopedia of Psychological Assessment*. Thousand Oaks, CA: Sage, pp. 960–64.

and Leung, K. 1997. *Methods and Data Analysis for Cross-Cultural Research*. Newbury Park, CA: Sage.

and Poortinga, Y. H. 1997. "Towards an integrated analysis of bias in cross-cultural assessment", *European Journal of Psychological Assessment* 13: 29–37.

and Poortinga, Y. H. 2002. "Structural equivalence in multilevel research", *Journal of Cross-Cultural Psychology* 33: 141–156.

and Tanaka-Matsumi, J. 2008. "Cross-cultural research methods". in D. McKay (ed.), *Handbook of Research Methods in Abnormal and Clinical Psychology*. Thousand Oaks, CA: Sage.

and Tanzer, N. K. 2004. "Bias and equivalence in cross-cultural assessment: an overview", *European Review of Applied Psychology* 54: 119–35.

van de Vliert, E., Huang, X., and Levine, R. V. 2004. "National wealth and thermal climate as predictors of motives for volunteer work", *Journal of Cross-Cultural Psychology* 35: 62–73.

Vandenberg, R. J. and Lance, C. A. 2000. "A review and synthesis of the measurement invariance literature: Suggestions, practices and recommendations", *Organizational Research Methods* 3: 4–70.

Wasti, S. A. 2002. "Affective and continuance commitment to the organization: test of an integrated model in the Turkish context", *International Journal of Intercultural Relations* 26: 525–50.

and Ondev, C. 2008. "Commitment across cultures: progress, pitfalls, and prepositions." in H. Klein, Becker, T., and J. P Meyer (eds.), *Commitment in Organizations: Accumulated Wisdom and New Directions*. Philadelphia, PA: Lawrence Erlbaum.

Welkenhuysen-Gybels, J., Billiet, J., and Cambré, B. 2003. "Adjustment for acquiescence in the assessment of the construct equivalence of Likert-type score items", *Journal of Cross-Cultural Psychology* 34: 702–22.

Werner, O. and Campbell, D. T. 1970. "Translating, working through interpreters, and the problem of decentering", in R. Naroll and R. Cohen (eds.), *A Handbook of Cultural Anthropology*. New York, NY: American Museum of Natural History, pp. 398–419.

Whiting, B. 1976. "The problem of the packaged variable", in K. Riegel and J. Meacham (eds.), *The Developing Individual in a Changing World*, vol. 1. The Hague: Mouton, 303–309.

Williams, K. Y. and O'Reilly, C. A. 1998. "Demography and diversity in organizations: a review of 40 years of research", in B. Staw and L. L. Cummings (eds.), *Research in Organisational Behaviour* 20: 77–140. Greenwich, CT: JAI Press.

Yammarimo, F. J. and Dansereau, F. 2002–2007. *Multilevel Issues in Organizational Behaviour and Processes*, vols. 1–7. Amsterdam: Elsevier.

and Markham, S. E. 1992. "On the application of within and between analysis: are absence and affect really group-based phenomena?" *Journal of Applied Psychology* 77: 168–76.

Yang, K. S. and Bond M. H. 1990. "Exploring implicit personality theories with indigenous or imported constructs: the Chinese case", *Journal of Personality and Social Psychology* 58: 1087–95.